Marginally discriminated: the role of outcome tests in European jurisdiction

Ilic, Dragan

Abstract: For decades, racial profiling has been subject of intense debate in US jurisdiction. Recently, outcome tests based on economic models have contributed to the legal discourse. However, it is not readily obvious if and to what extent they also pertain to European jurisdiction, where racial profiling has only as of late stirred up controversy. In a comprehensive examination of their basic building blocks, this paper illustrates why the these tests are not particularly suited for the European case. The models are tailored to identify racial prejudice but are unfit to provide evidence of statistical discrimination, reflecting their adaption to the current US legal approach. A simple alternative test remedies this shortcoming and manages to inform the European jurisdiction.

DOI: https://doi.org/10.1007/s10657-013-9409-9

Eur J Law Econ (2013) 36:271–294
DOI 10.1007/s10657-013-9409-9

Marginally discriminated: the role of outcome tests in European jurisdiction

Dragan Ilić

Received: 1 September 2010 / Accepted: 8 July 2013 / Published online: 21 August 2013

Abstract For decades, racial profiling has been subject of intense debate in US jurisdiction. Recently, outcome tests based on economic models have contributed to the legal discourse. However, it is not readily obvious if and to what extent they also pertain to European jurisdiction, where racial profiling has only as of late stirred up controversy. In a comprehensive examination of their basic building blocks, this paper illustrates why the these tests are not particularly suited for the European case. The models are tailored to identify racial prejudice but are unfit to provide evidence of statistical discrimination, reflecting their adaption to the current US legal approach. A simple alternative test remedies this shortcoming and manages to inform the European jurisdiction.

Keywords Racial profiling · Outcome test · Discrimination · Europe

JEL Classification J71 · K42

1 Introduction

In the aftermath of the terrorist attacks of September 2001, racial profiling, a controversial policing practice that had hitherto been predominantly subject of debate in the US, began to gain ground in Europe. Racial profiling describes the use of ethnicity or race as one of the criteria of law enforcement officers in deciding

D. Ilić (✉)
Center for Corporate Responsibility and Sustainability, University of Zurich, Zähringerstrasse 24, 8001 Zurich, Switzerland
e-mail: dragan.ilic@ccrs.uzh.ch

D. Ilić
Faculty of Business and Economics, University of Basel, Peter Merian-Weg 6, 4002 Basel, Switzerland

whether to (preemptively) stop or search suspects. This method was overtly suggested even by leading figures after the subsequent terrorist attacks in Europe. As Ian Johnston, the Chief Constable of British Transport Police put it after the attacks in London in July 2005: "We should not waste time searching old white ladies. It is going to be disproportionate. It is going to be young men, not exclusively, but it may be disproportionate when it comes to ethnic groups." (Dodd 2005) The introduction of racial profiling in Europe has stirred up major public controversy but, in contrast to the US, it has not yet been extensively discussed in the judicial realm (Baker 2007).

Assessing the motivation behind racial profiling gives rise to intricate issues. Some critics argue that the discretionary nature of police searches gives leeway for prejudice and thus racial animus, enabling racist officers to indulge in their malevolent preferences. This argument goes beyond the mere consideration of ethnicity in the decision-making process. To a racist officer, searching individuals of a despised ethnicity is an end to itself. In contrast, the potential consideration of suspect ethnicity by an unprejudiced officer would serve as means to optimizing the odds of catching actual perpetrators, notably the proclaimed goal of any form of profiling. Both varieties of discrimination—known in the economic literature as taste-based and statistical, respectively—burden the affected ethnicities with a greater likelihood for stop or search *because* of their ethnicity. Yet a greater likelihood in the empirical data does not imply that ethnicity played a role in the profile. For it is possible that ethnicity merely correlates with other observable characteristics indicative of criminal intent. A policy that bases its search decisions solely on these characteristics will nevertheless end up having a disparate impact in form of differing stop or search rates. This poses a problem in the interpretation of empirical outcome data with respect to the underlying motive. All three explanations—animus, efficiency, and correlation—are consistent with disproportionate impact. The cause of the disparity, however, is not readily obvious from the available data.

Disentangling the motives is of high importance from a legal standpoint. The US and European jurisdictions differ substantially in their understanding of justified discrimination in police searches. In the US, discrimination law has no issues with disparate impact. Instead, the plaintiff has to establish an intent to discriminate, which is defined as prejudice that serves no policing purpose (Persico and Todd 2008). In Europe, on the other hand, proving mere disparate impact is sufficient for a case to hold up in court (Baker and Phillipson 2011). A judge applying European law as defined in the European Convention on Human Rights does not require evidence of malevolent intention to establish unlawful policing. Even so, the motive and hence actual cause behind an outcome disparity is also significant in European jurisdiction for it weighs up matters of proportionality; the question whether the disparate impact can be traded off in light of a compelling state interest.

What has economics to say about the causes of disparate impacts? Two methodologically distinct branches have evolved in the literature to infer motives from empirical outcome data. The first branch deals purely statistically with the question whether police officers treat similarly suspicious individuals differently because of their ethnicity. Typically, a multiple regression conditions the dependent

variable (say, the probability of a search) to a list of independent variables that are potentially informative of engagement in criminal activity (such as behavior, location, or demeanor of the suspect). The difficulty that arises from this approach is that ex post, the researcher rarely gains insight into all the variables that make up the suspicion signal. This classic statistical difficulty is known as the omitted variables problem. Even if one had access to all this information, it is arguable whether it can be quantified appropriately for statistical inference.

The second, more recent branch of the racial profiling literature circumvents these problems by tackling the question from the other end. In focusing on the outcome of the decision-making, a researcher can make inferences about the level of "probable cause" that the decision-maker has required during the prior assessment. For instance, if searches against minority groups yield lower success rates than against other groups, one would deduce that the police unjustifiably apply a lower threshold of suspicion when deciding to search minorities. Crucially, the omitted variables problem does not arise in this analysis. All the relevant variables were already taken into account by the decision-maker. This information is reflected in a single statistic: the outcome. If the decision-maker only cares about efficiency and thus aims for the optimal outcome, there should be no disparities along ethnicity or gender in the outcome data. This method of inferring prejudice is appropriately called an outcome test.

In the context of racial profiling, modeled adaptions of outcome tests to determine police prejudice in motor vehicle searches have emerged in the US during the last decade. They have become increasingly popular in the economics literature and are of practical relevance for the judicial realm (Knowles et al. 2001; Anwar and Fang 2006). The theoretical foundations of these modeled outcome tests address some of the drawbacks of the traditional outcome test application. Obviously, the models have been adapted to the specific context of the US law. In particular, they are apt at distinguishing racial bias (the illicit intent to discriminate) from legal statistical discrimination. While disentangling these two motives is valuable to the European jurisdiction as well, it is not clear what the models have to say about the existence of statistical discrimination itself. However, in European jurisdiction it is central know whether any disparate impact is the cause of deliberate statistical discrimination or whether it originates from a race-neutral policy. This difference affects the justification of disparate impact.

The literature on modeled outcome tests has neglected this crucial difference in European jurisdiction. This paper elaborates on the exact role of statistical discrimination in the existing models and shows which lessons can be drawn for the judicial discourse in Europe. The paper makes no contention about the ethical scope of racial profiling (Risse and Zeckhauser 2004; Lever 2005). It also ignores any detrimental effects towards social cohesion and stigmatization (Loury and Coate 1993; Loury 2002). Neither will I dwell on questions about the validity of rational choice assumptions in law and economics or well-known caveats of cognitive bias in behavioral economics (although rational choice is discussed critically). Instead, I illustrate why and to what extent the existing literature is relevant for European racial profiling litigation. I contend that while the existing modeled outcome tests remain somewhat useful for the European approach, they are not suitable to address

proportionality and thus the justification of disparate impact. I conclude by proposing a simple alternative way to answer this question.

The paper proceeds with a theoretically based motivation for the emergence of (modeled) outcome tests and starts with a brief introduction to the basic building blocks of the economics of discrimination. These theoretical foundations are described in the next section. Section 3 builds on this formal basis and exposits the two dominant economic models on prejudice in motor vehicle searches. A particular focus is given to the role of rational choice and statistical discrimination. Section 4 lays out the central differences in the US and the European jurisdiction on racial profiling and highlights the role of the models for Europe. Section 5 concludes.

2 Theoretical foundations of the outcome test

Economic theory distinguishes between two forms of discrimination. Taste-based discrimination corresponds to malevolent intent, unjustified prejudice so to speak. This form of discrimination is an end to itself as the discriminating behavior provides actual utility to the offender. Statistical discrimination, on the other hand, denotes the instrumental use of race (or, say, gender). If race improves an assessment of some form of productivity, statistical discrimination contends that rational and unprejudiced decision-makers should take group affiliation into account. This section briefly outlines the theories of taste-based and statistical discrimination and explains how outcome tests incorporate both notions.

2.1 Taste-based discrimination

The economic study of malevolent discrimination has its roots in the works of Becker (1957). In his model, discrimination is part of an individual's utility function. Following Altonji and Blank (1999), Becker splits his analysis into three parts: employer discrimination, employee discrimination, and consumer discrimination. To get an idea of the basic principle, it suffices to present a very general formulation of the model. The key aspect in Becker's framework is the proposition that a so-called "taste for discrimination" directly affects the utility in case of association with the group against which the taste is directed. For instance, an employer suffers a loss of utility by employing some person X, an employee suffers a loss of utility by having to work with person X, or consumers experience a utility loss when buying from person X. This proposition extends the traditional profit or utility function. In the employer-employee example, the employer—or rather, the firm—maximizes the following function:

$$U = pF(N_w + N_b) - \omega_w N_w - \omega_b N_b - dN_b$$

Note that in contrast to conventional profit maximizing behavior, firms are assumed to maximize utility. The production function is given by F and the exogenous price level is given by p. N_r with $r \in (w, b)$ denotes the number of employed white (w) and black (b) workers at wages ω_r, respectively. So far, these terms coincide

with a conventional profit function. Discrimination against black workers extends this function and operates through the coefficient $d > 0$. This so-called discrimination coefficient measures the cost of employing N_b black workers.

Discriminating firms may differ in their intensity of discrimination, such that d is distributed according to the cumulative density function G with mean $\overline{d}$. The firms' malevolent preferences increase the price of hiring a black worker to $\omega_b + d$. In essence, a discriminating firm behaves *as if* the wage of a black worker was $\omega_b + d$. Thus, a firm will only hire black workers if their price is low enough in comparison to white workers. A firm maximizes its utility by setting the marginal value of a worker equal to his or her marginal cost.

What determines the observed wage gap in the market? One might be inclined to think that $\overline{d}$, the average degree of discrimination among the firms, determines the market wage of black workers. Instead, it turns out that the marginal firm which is indifferent between hiring a white or black worker is decisive for the wage gap. Market wages for black workers are determined by the firm with the *highest* taste for discrimination among the firms that are willing to employ black workers. Note that all firms with d lower than the marginal firm are at an advantage, while all firms above the marginal firm would make a loss from employing blacks.

Becker's model predicts the eradication of discrimination in many market settings through the force of competition. So ironically, the model "predicts the absence of the phenomenon it was designed to explain" (Arrow 1972, p. 192). But tastes alone seem to be unable to explain ongoing observable differences in central economic variables along race and gender. While US data do show some convergence in the decades after the legal ban of overt discrimination with the Civil Rights Act of 1964, empirical evidence also suggests that the trends stagnated in the 1980s and 1990s. Moreover, even today taste-based discrimination is seemingly present in expensive markets such as car sales and housing (Yinger 1998). This indicates that competitive pressure on its own is not likely to equalize differences in market outcome. Against this backdrop, in the 1970s models of imperfect information extended the economic theory of discrimination. The next subsection exposits the basic principles of statistical discrimination.

Two basic principles of the taste-based framework are still reflected in today's models. First, racial animus affects the utility of the discriminating individual. Costly association is one representation of this effect. Other approaches model a benefit from discrimination, deriving from the mistreatment of despised groups. Second, the cause of disparate impact should be evaluated at the margin, not on average. In Becker's model, it is the marginal prejudiced firm, and not the average one, that determines the level of disparate outcome in the market.

2.2 Statistical discrimination

The seminal work by Phelps (1972) reshaped the economic analysis of discrimination. Instead of harboring malevolent tastes, decision-makers use race or gender as a *signal* to improve an assessment under uncertainty. Phelps takes the example of

noisy signals of productivity that are transmitted from employees (agents) to employers (principals). The optimal solution to this signal extraction problem is statistical discrimination. This form of discrimination can explain inequalities in outcomes in the absence of prejudice. In what follows, I briefly discuss the statistical discrimination models laid out in Aigner and Cain (1977).

Assume that there are agents with a true skill level q which principals can only observe indirectly via an indicator y, some form of aptitude test. While the indicator measures the true skill level on average, there is uncertainty in the assessment because the relationship is diluted with an error term:

$$y = q + u$$

The distribution of the error term u is i.i.d. and follows $\sim N\left(0, \sigma_u^2\right)$. The true skill level q is i.i.d. according to $\sim N\left(\alpha, \sigma_q^2\right)$. These distributions are public information. Principals are interested in an estimator of q given y. They do best to form an "educated guess" based on all the available information. The expectation $\widehat{q}$ is formally given by

$$\widehat{q} = E(q|y) = (1 - \beta)\alpha + \beta y. \tag{1}$$

The expectation of the true skill level is the weighted sum of a group term (with an average true skill level α) and an individual term (with an indicator level y). The parameter β scales the individual signal y . A high β means that the signal is a good indicator of the true skill level.

In the simple job market environment described in Aigner and Cain, the expectation of the true skill level $\widehat{q}$ is assumed to be directly linked to the agent's wage. Equation (1) states that $\widehat{q}$ is a function of y. In particular, the expectation of the true skill level $\widehat{q}$ depends on (a) the average skill level α, (b) the variance of the true skill level q, and (c) the variance of the error term u which dilutes the individual signal. These two variances enter Eq. 1 through β, which is defined as:

$$\beta = \frac{\sigma_q^2}{\sigma_q^2 + \sigma_u^2} \tag{2}$$

Equation (2) illustrates to what extent information about the group affects optimal inference of individual signals. Consider two groups of workers, black and white, which differ only in their variance of true skills, σ_q^2. Assume that the true skills among black workers vary more than among white ones. The average skill as well as the extraction quality via the indicator y is equal for both black and white workers. It then follows from Eq. (2) that β is higher for black workers. In other words, the indicator y is a more reliable signal of the true skill for black workers if true skill varies more in their population.

What does this entail for the wages paid by the employers? In an unprejudiced environment, all workers are paid based on their *expected* skill level. Any worker with an indicator that coincides with the average of the true skill α is going to earn exactly according to his or her true skill. However, black workers with indicators above α are paid more than white workers with equal indicators. Because y is a more reliable signal for the true skill of black workers, employers can be quite confident

that black workers with an above average indicator are actually better than white workers with the same indicator. On the other hand, below average indicators y result in lower wages for black workers because employers infer that they are more likely to have less skills.

In this example, we will find relatively lower wages for black workers in low productivity jobs and relatively higher wages for black workers in high productivity jobs. If we flip the situation and assume the indicator of white workers to be more informative about their true skill, we would observe the opposite. So in the absence of prejudice, the Phelps model can explain unequal compensation for black and white workers with the same indicator y.

Differences in average group abilities also give rise to unequal compensation of workers with equal indicators y. Assume that both the variances of the true skill levels and the error terms are equal for black and white workers. The quality of the predictor as reflected by β is therefore equivalent. Now let the average true skill α be higher for white workers. In this setting, black workers consistently earn a fixed amount less than white workers at every indicator level y.

Statistical discrimination is a indispensable ingredient in modern economic models of discrimination (for a comprehensive survey, see Fang and Moro 2011). The models have become increasingly sophisticated, but the basic principle of optimal signal extraction remains: If individual signals are imperfect, observable group affiliation can be useful to improve the principal's assessment.

2.3 Outcome tests

Outcome tests are an established alternative to traditional statistical analyses to identify taste-based discrimination in the marketplace. The method relies on the premises that animus is costly and that decision-makers make full use of available information. In other words, the method draws from the principles of taste-based and statistical discrimination. Fittingly, it was Becker (1993) who made this idea popular. Becker insinuated that prejudiced banks could be identified by comparing mortgage default rates of black and white customers.

To illustrate, let a bank rank their mortgage applicants by ascending credit ratings, starting with the ones that are most likely to default. The bank chooses a certain threshold of creditworthiness above which it deems the risk of default low enough. Only applicants above the threshold are granted a mortgage. If the bank applies the same standard to both black and white applicants, the outcome test claims that the probabilities of default should be equal. The rejection of relatively good risks results in lower average default rates. So if black lenders display a lower default rate, the bank is concluded to have intentionally applied a more stringent credit standard and therefore to be racially prejudiced. In a nutshell, outcome tests infer if a principal requires better outcomes for certain groups when making a decision.

Outcome tests can be applied to a myriad of settings in which a principal demands productive outcomes from an agent. The existing literature covers subjects such as lending decisions, bail bond settings, paper citation rates, organ transplantation, and, of course, police searches. I briefly discuss the underlying ideas in turn.

Bail bonds create incentives for suspects to appear in court for an upcoming trial. The price of the bail bond is set by the judge, who weighs up the benefit of the suspect not having to stay incarcerated against the risk of the suspect not appearing for trial. If black suspects appear more frequently for given bail bond prices, the judge is assumed to have set the bail too high because of prejudice. Ayres and Waldfogel (1994) analyze appearance rates for a sample of US trials and conclude an unjustifiably higher standard required from black suspects.

Smart et al. (1996) test whether publications of female authors in refereed economic journals display deviating citation rates. If their papers yield higher citation rates, female authors would seem to have satisfied tougher requirements from referees and editors. The study finds no gender bias.

Health economics also provides fertile ground for the application of outcome tests. Ayres (2005) argues that if minorities end up having longer survival times after a kidney transplantation, they must have unjustifiably satisfied higher standards of health prospect in order to receive the transplant.

Finally, Ayres (2001, 2002) describes the use of outcome tests to identify racial prejudice in motor vehicle searches. Knowles et al. (2001) put forth an according model which gives rise to testable predictions. In an nutshell, in their model the police are deemed unprejudiced if the probabilities of a successful search do not differ between black and white motorists. On the other hand, if black motorists were victims of taste based discrimination, the police would apply a lower suspicion threshold when deciding whether to search them. This would yield a lower probability of uncovering engagement in criminal activity among black drivers. Anwar and Fang (2006) address some drawbacks of the model and present an alternative test. For reasons I will discuss in a bit, both these outcome tests for racial prejudice in motor vehicle searches are based on models with specific economic assumptions. The models are exposited in Sect. 4.

The advantage of the outcome test is that it is not susceptible to omitted variable bias. This bias can occur if the statistician has less information available than the decision-maker. Consider a potential race or gender bias against applicants during job interviews. To identify this bias by traditional regression analysis, data are gathered ex post to calculate the likelihood of getting a job depending on the applicant's characteristics. For the sake of the argument, let us assume that applicants from Harvard are particularly apt for the job in question, and the employer knows this. Let "Harvard graduate" be a feature that occurs with higher propensity among white males. If this information is not available to the researcher, one could be mistakenly be lead to believe that white males receive preferential treatment. Outcome tests avoid such confoundings. Because the decision-maker has plausibly considered any variable that affects the desired outcome, the statistician has indirectly full access to all this information via the outcome data.

To sum up, an unbiased employer should only care about productivity. If ex post, there is an observable difference in productivity along race or gender, the employer required a higher quality from the discriminated group. In terms of Becker's model, the quality gap is a measure of the employer's malignant utility. This preference can be modeled in various ways. One can argue that employers suffer utility losses from giving jobs to applicants they feel contempt for. At some point however, the

productivity of a highly qualified minority applicant begins to make up for these costs such that the employers cannot afford to decline jobs to all minorities. An alternative modeling of discriminatory preference is more sadistic: A malevolent employer might actually enjoy treating the disadvantaged group unfairly. Whatever the reason, an unfair decision will lead to differences in productivity ex post. In addition to this utility concept, outcome tests rely on the premise of statistical discrimination. The principal is assumed to make best use of all available information. If group affiliation improves the assessment, it will be taken into account.

2.4 Drawbacks of outcome tests

In light of limited information, outcome tests seem like an intriguing approach to assess the motives behind a decision-making. However, the method has its drawbacks. There are two main issues. The first one relates to the identification of statistical discrimination and pertains directly to the significance of outcome tests in European jurisdiction. The second issue is known as infra-marginality and questions the applicability of the test under certain conditions. Let us first discuss the role of statistical discrimination.

Outcome tests assume that all the information on which the principals base their decisions is fully exploited in terms of efficiency. This is a crucial assumption. It means that any applicant signal that improves the assessment of the productivity is taken into account. Section 2.2 has shown that group affiliation is of potential use when making an educated guess and forming expectations based on noisy signals. So when faced with uncertainty, in outcome tests the principal is bound to discriminate statistically if visible group affiliation is useful. More precisely, if the principal was able to improve the assessment by incorporating statistical discrimination, outcome tests implicitly require him to have done so by assumption. From a rational point of view, this assumption does not seem not so far-fetched. If an unbiased principal realizes group differences in desirable outcomes, she will update her assessment in light of the costly consequences of her decisions. The use of statistical discrimination avoids losing good risks.

But this assumption becomes a practical issue when statistical discrimination is not applicable. Assume that the researcher ex post has access to some information that the principal did not have. If this information had improved the productivity assessment, outcome tests can lead to false conclusions, in particular when observability for the principal changes. Table 1 adapts an example from Persico (2009) which describes drug trafficking and police searches. For now, assume that

Table 1 Observability constraints in outcome tests

	Black motorist	White motorist
Red car	50 %/50	50 %/50
Blue car	0 %/70	60 %/70

there is a homogenous distribution of criminal likelihood for each category. The argument readily extends to heterogeneous distributions that invoke suspicion thresholds.

Consider a police officer who has no taste for discrimination but wants to maximize the probability of uncovering criminal engagement during a motor vehicle search. The officer has a budget of 100 searches. Two characteristics affect the allocation of the searches: the color of the cars, and potentially motorist race. Table 1 shows the known and exogenous probabilities of criminal activity for each car color/motorist race combination as well as their absolute frequencies.

In a first step, let statistical discrimination by race not be applicable, either because the race of the motorist is not visible or simply illicit in the search profile. Car color, on the other hand, can always be exploited for inference. In this case, the officer does best to search red cars only. In contrast to searching blue cars, this yields 50 hits compared to 42. Both the search success rates against black and white motorists are 50 %. A researcher conducting an outcome test on motorist race would therefore conclude the absence of racial animus.

If however the race of the motorist can be taken into account to improve the assessment, the officer would search all white motorists driving a blue car and devote the rest of the budget to motorists driving red cars, no matter black or white. This strategy yields 57 hits. If the 30 searches against red cars are randomly distributed between black and white motorists, we expect 15 of each to be searched on average. An outcome test on motorist race now concludes racial animus: The search success rate against black motorists is $7.5/15 = 50$ %, whereas white motorists are searched successfully at a rate of $(7.5 + 42)/(15 + 70) = 58.24$ %.

The color-blind case in the example highlights that an outcome test for prejudice might not be applicable if the researcher has access to more information than the principal. More specifically, when statistical discrimination was not employed despite its potential to improve the assessment, outcome tests based on these very characteristics can be misleading. By the same token, outcome tests are not applicable when statistical discrimination is used reluctantly or is forbidden by law. Ayres (2005) makes a similar argument in the case of kidney transplants. The effectiveness of a mandated color-blind or gender-neutral policy to eliminate prejudice can therefore not be assessed with outcome tests. The example also stresses that in general, animus cannot be concluded from changes in treatment if the characteristic in question becomes observable. For example, a rise in job offers to female musicians after the introduction of blind auditions has to be interpreted carefully (Persico 2009). Changes in treatment could be triggered by statistical discrimination.

The second issue associated with outcome tests is dubbed the infra-marginality problem. Outcome tests usually assess average outcome data. But conclusive inferences about the principal's motives need to be assessed at the margin. Yinger (1996) explains the problem in the mortgage context as follows.

Assume that banks rank their mortgage applicants by creditworthiness. All applicants must fulfill a certain degree of some objective merit criteria. The applicant who just qualifies for a mortgage is the marginal applicant. If the bank is unbiased, the probabilities of default of the *marginal* black and white applicants are

equal. However, the *distribution* of the probabilities of the good risks above the threshold are not necessarily equal. Consequently, the average default rates among the ones that make the cut will typically differ and are not indicative of marginal decision-making.

Heckman (1998) describes a vivid sports analogy for this inference problem. Consider black and white high jumpers who have the same athletic ability to jump but use different techniques. Although the mean height they achieve is equal, the variance in heights is lower in the population of white jumpers. In other words, black jumpers are more likely to reach both lower and higher heights. Unbiased treatment implies that the bar is set at the same height for everyone. But if the bar is set rather high, relatively more black jumpers will excel. On the other hand, if the bar is set fairly low, black jumpers will be more likely to fail than white jumpers. The example goes to show that differences in average failure rates cannot be ascribed to unfair treatment if the groups do not share the same qualification pool. Indeed, an outcome test could mistakenly indicate the absence of racial bias when in fact it is present. Worse still, an outcome test could mistakenly indicate racial bias against the actually privileged group. Infra-marginality, then, poses a serious problem for the applicability of outcome tests. If there is no good reason to assume that the different groups share the same quality distribution, average outcome data is of limited use for the inference of animus.

The infra-marginality problem translates to the context of motor vehicle searches. In general, racial animus cannot be ruled out on the basis of equal average search success rates. To illustrate, line up all motorists according to their probability of carrying contraband, based on a police officer's perception. Consider an unbiased officer that will only search motorists that are above a specific probability of guilt threshold that applies to both black and white motorists. If the distributions with respect to these probabilities differ such that black motorists above the threshold are, on average, less likely to be guilty, the outcome test will falsely conclude racial animus. And to put the above generic pitfall into context, an outcome test might even indicate racial bias against white motorists when in reality the officer is biased and demands less scrutiny from black motorists.

Motor vehicle searches depict situations with dichotomous outcomes. The police are either successful or not. Mortgage applications follow the same pattern. Applicants either default or not. However, infra-marginality is less of a problem in a non-dichotomous setting. Take the examples of article citations or bail bonds. Articles receive a countable number of citations and bail bonds are set at a specific amount for each case. In both settings, the according distributions can be assessed more closely and marginal decision-making becomes more apparent due to the particular outcome for each entity.

To round up, outcome tests rely on two crucial assumptions. First, the decision-maker is assumed to make use of statistical discrimination if it improves the assessment. The inferences based on outcome tests have nothing to say about whether statistical discrimination was *actually* applied during the decision-making. Either way, observable differences in outcomes are attributable to non-productive motives, namely animus (which is traded off for productivity). Conversely, this implies that if beneficial statistical discrimination was not applicable during the

decision-making, outcome tests are not informative of animus. The second crucial assumption outcome tests rely on is that they require the groups in question to share the same pool of qualities. Inference based on average outcome data is thus not appropriate if the groups display differing distributions of outcome-relevant qualities.

Economic theory rarely has issues with the first assumption. On the contrary, full use of available information is embedded in the modeling core of classic economic behavior. The second assumption raises an application problem, however. This explains the recent rise of modeled outcome tests which are built to tackle infra-marginality in effort to provide eligible inferences. On that note, Anwar and Fang (2013) study bounceback rates of emergency room visits to assess racial prejudice in health care. Alesina and La Ferrara (2011) provide a rank order test for racial prejudice based on judicial errors in lower courts. Anwar and Fang (2012) model the decision problem of a parole board and put forth a test for racial prejudice in parole decisions. Finally, since September 2001 airport searches have become subject of heated debates. Persico and Todd (2005) show the trade-offs involved in a model of passenger profiling.

To assess racial prejudice in motor vehicle searches, Knowles et al. (2001) (henceforth KPT) put forth a seminal model that relies on rational choice on the part of both the motorists and the police. A subsequent model by Anwar and Fang (2006) relaxes the assumption of rational choice and instead exploits data on officer race. The next section exposits how both models address infra-marginality and highlights the role of taste-based and statistical discrimination.[1] The analysis will make clear why the testable implications of the models remain of limited use for racial profiling litigation under European jurisdiction.

3 Two models of racial profiling in motor vehicle searches

3.1 Rational choice: the KPT framework

There is a continuum of homogenous police officers controlling motorists with observable race $r \in \{B, W\}$. The police aims to uncover engagement in criminal activity, say, motorists carrying contraband. Let c be a one-dimensional variable—partially or fully unobservable by the statistician—which combines all variables other than race that affect an officer's search decision. The cumulative distributions of c among black and white motorists are given by $F(c|B)$ and $F(c|W)$. Police officers maximize the probability of finding contraband minus their search cost. The benefit of an arrest is one and the marginal cost of searching an r_m motorist is $t_r \in (0, 1)$. Officers are said to exhibit a taste for discrimination if these costs depend on the race of the motorist. A successful search, i.e. finding contraband, is indicated by the event G. For simplicity, assume that guilty motorists are always uncovered if they undergo a search.

[1] Some passages in Sect. 3 and, to a lesser extent, Sect. 4 recapitulate the descriptions in Ilić (2013).

In this game, police officers make a decision whether to search based on the observable information (c, r). Motorists, on the other hand, decide whether to carry contraband. Consider first the motorist's decision. The search probability is the only endogenous factor in the decision whether to carry contraband.[2] Should they decide not to carry contraband, their payoff is zero regardless of being searched or not. Should they decide to carry, their payoff is $-j(c, r) < 0$ if they are searched and thus uncovered. If they get away with carrying contraband, their benefit amounts to $v(c, r) > 0$. The search probability of a motorist of type (c, r) is denoted by $\gamma(c, r)$. Taking all this into account, the expected payoff of the motorist becomes

$$\gamma(c,r)\{-j(c,r)\} + \{1-\gamma(c,r)\}v(c,r).$$

Carrying contraband is profitable if the expected payoff from doing so is positive. Conversely, the motorist does not carry contraband if the expected payoff is negative.

Now turn to the decision problem of the police. Based on the guilt probability $P(G|c, r)$ the officer decides at what rate to search each motorist group. Police officers maximize their payoff by setting their search rates $\gamma(c, r)$ as to maximize

$$\max_{\gamma(c,W),\gamma(c,B)} \sum_{r=W,B} \{P(G|c,r) - t_r\}\gamma(c,r)f(c|r)dc.$$

If the term inside the curly brackets—the officer's benefit of a successful search minus the search cost—is greater than zero, the officer will search the according type (c, r) with probability one (and vice versa).

The situation implies a mixed Nash equilibrium in which both motorists and police officers randomize their strategies. The indifference condition for the motorist yields the equilibrium search intensity

$$\gamma^*(c,r) = \frac{v(c,r)}{v(c,r)+j(c,r)} \tag{3}$$

The indifference condition for the police officers determines the guilt probability (or search success rate)

$$P^*(G|c,r) = t_r, \forall c, r$$

if the police officers do not exhibit a taste for discrimination. Thus, in an unprejudiced equilibrium, black and white motorists carry contraband with equal probability. If the probabilities between the groups differed, the officers would (completely) devote themselves to the group with the higher propensity. This in turn would affect the incentives in the neglected motorist group. A pure search strategy can therefore not be optimal.

Racial animus is modeled by search costs that depend on motorist race, $t_B \neq t_W$. This can readily be tested with outcome data. Should the guilt probability (or rather, the search success rates) of searched motorists vary by race, the police officers trade off some of the benefit of an arrest with the benefit derived from racial animus.

[2] In the economic approach to crime pioneered by Becker (1968), two factors determine the decision of engaging in criminal activity: Measure of the punishment, and probability of getting caught. Whereas the former is considered fix in the KPT model, the uncovering probability is determined endogenously.

In terms of the model, lower search costs against a discriminated group lead to "oversearching". The increased deterrence drives down the search success rate, precisely the effect that can be empirically tested for.

A central insight of this model is the unbiased explanation for higher search rates against one group. This happens if either the expected value of carrying contraband or the cost of being found guilty varies by motorist race. A higher incentive of carrying contraband moves in lockstep with a stronger deterrence in form of a higher search rate, dictated by the equilibrium condition of equal search success rates.

KPT test their model on 1,590 observations of motor vehicle searches on a highway stretch in Maryland, USA from 1995 to 1999. After the first lawsuit filed by The American Civil Liberties Union of Maryland in 1993, the police started collecting systematic information on motor vehicle searches. KPT's empirical test compares the search success rates against black and white motorists and cannot reject equality, suggesting that police officers in the Maryland data do not exhibit racial animus against black motorists. Their results do however imply that Hispanic motorists suffer from prejudice.

The KPT model incorporates the two classic economic branches of taste-based and statistical discrimination discussed in Sect. 2. Willingness to pay for malevolent tastes is modeled by trading off the benefit of higher search success rates. The lower payoff from searching the discriminated group is offset by the benefit of mistreatment. This gap is equivalent to the wage gap between black and white workers in Becker's employer discrimination model. The significance of marginal decision-making is reflected in KPT as well. The system resides in equilibrium through a marginal trade-off in search success rates. But in contrast to Becker's employer discrimination model, racially biased police officers are not driven out of the market. They bear the foregone profits in terms of lower hit rates themselves in a non-competitive market.

In addition to taste-based discrimination KPT's model incorporates the notion of statistical discrimination. The police observe the individual signal c as well as group identity r and do best to exploit this information fully. But do we know if the police actually make use of group identity? For example, assume that in an unbiased equilibrium black motorists are searched at a higher rate. Does this disparate impact stem from statistical discrimination? It turns out that the KPT model cannot answer this question.

On the one hand, it is possible that the police do use motorist race to make inferences about unobservable characteristics that are correlated with crime. For example, r might constitute a proxy for crime if lower earnings are associated with a higher propensity to carry contraband. Assume that the black population earns less on average. In this case, black motorists with the same observable characteristics c as white motorists are searched at a higher rate if the proxy carries additional information indicative of crime (KPT, p. 212). In other words, white motorists get to send a higher signal of suspicion c before they are deemed suspicious enough to search. This relates to Aigner and Cain's classical model of statistical discrimination in Sect. 2.2, where employers are compensated differently despite the same productivity indicator.

On the other hand, the differences in search rates may solely arise from differences in the distribution of c, the characteristics indicative of crime that the police actually observe. These distributions might differ by race such that the individual signals of suspicion prompt the police to search black motorists at a higher rate. Race then merely correlates with the signals and does nothing to improve the assessment. In other words, the police ignore race.

The latter explanation does not constitute racial profiling because race is not used to make inferences about engagement in criminal activity. The first explanation does fit the definition of racial profiling for optimal signal extraction. However, both explanations are consistent with higher search rates, and it may be the case that both explanations apply. The KPT model has nothing to say about the actual cause. The inability to detect statistical discrimination is a disadvantage of the KPT model, and unavoidably so because it is based on the general outcome test methodology. In return, the KPT test inherits the primary advantage of outcome tests and does not suffer from omitted variable bias. The police officers make use of all observed information, information that is eventually reflected in the search success rates.

In contrast to the issue of statistical discrimination, the KPT test successfully deals with the second issue in outcome tests. Like all outcome tests, KPT lends itself to empirical application because of its modest information requirement. Section 2.4 has shown that traditional outcome tests have problems in interpreting average outcome data because group distributions in guilt probabilities may differ. The KPT test, however, is not susceptible to this infra-marginality problem. Because of the strategic interaction in the model, the (unprejudiced) equilibrium predicts $\pi(c, r) = t_r$ for all c, so all motorists are equally likely to carry contraband. The marginal and the average search success rate are one and the same. This feature renders average outcome data informative and allows for inferences about racial animus.

The follow-up literature has raised several objections against the KPT model. Let us first evaluate some that do the model no harm. The equilibrium prediction of all motorists randomizing on the decision to carry contraband does not seem sensible. KPT show that the model can be alternatively specified with a random variable relating to the motorists' utility from carrying. In said specification, motorists never randomize. But since the random variable is private information, the police cannot condition their search decision on the random utility. This alternative specification is "observationally equivalent" (KPT, p. 214) to the discussed model. Other critics question the assumption of strategic criminal decision-making of the entire motorist population. However, this assumption can be relaxed, making for a more realistic description. The model prediction does not change if we assume that only a minute (albeit racially equal) percentage of motorists would actually ponder criminal engagement by virtue of rational choice (Persico 2002, p. 1484).

Several papers extend and generalize the KPT model. Dharmapala and Ross (2004) constrain the assumption of perfect visibility of all motorists. In their model some motorists will always carry contraband, and the police will always want to search them when spotted. Antonovics and Knight (2004) introduce heterogeneous utility distributions for both motorists and police officers, an idea further embraced in Persico and Todd (2006).

Of all the objectionable assumptions in the KPT model, rational choice among motorists seems the most striking one. The model assumes that all *searched* motorists were well aware of the (exact) probability of a successful search when commuting on the motorway and behave accordingly. How sensible is this assumption? In a back-of-the-envelope estimation, Close and Mason (2007) guess that the actual probability of a random motorist in Florida being searched is 3 in 1 million. This greatly qualifies the deterrence effect that policing imposes, especially given the fact that drivers who abide by the traffic law are highly unlikely to be stopped in the first place. It is fair to assume that determined criminals will take heed of any behavior that might raise the search risk. However, without the strategic reasoning on the part of the motorists, the infra-marginality problem recurs, rendering inferences based on average outcome data invalid (KPT, p. 212).

The infra-marginality problem could also crop up for other reasons. Anwar and Fang (2006) (henceforth AF) stress that the KPT model fails if motorist behavior during a stop is indicative of criminal engagement. AF raise yet another issue. To date, all models had considered the police force as one homogenous entity. But if we allow black and white police officers to exhibit racial animus against white and black motorists, respectively, analyzing the aggregated police force data could lead to mistaken conclusions. AF propose an alternative model that takes these issues into account. What is more, their test for prejudice does not require the motorists to make a rational choice.

3.2 A diverse police: the AF model

Let there be a continuum of police officers and motorists of race r_p and $r_m \in \{B, W\}$, respectively. The police stop and decide whether to search motorists. A fixed fraction of the motorists is engaged in criminal behavior: $\pi^{r_m} \in (0, 1)$.[3] When stopped, motorists emit a signal $\theta \in [0, 1]$, a one-dimensional index capturing all relevant characteristics informative of criminal activity. The signal contains some uncertainty, but the police know that it is distributed according to the continuous probability density function $f_g^{r_m}(\cdot)$ if the motorist of race r_m is engaged in criminal behavior. If innocent, the index is drawn from $f_n^{r_m}(\cdot)$. It is not out of the question that motorists emit even high signals of guilt even though they are actually not guilty. The likelihood of this happening follows an intuitive formal condition. The two densities satisfy the strict monotone likelihood ratio property, such that $f_g^{r_m}(\theta)/f_n^{r_m}(\theta)$ is strictly increasing in θ. A higher θ therefore accurately indicates a higher guilt probability.

An r_p officer bears the marginal search cost $t(r_m, r_p)$. Note that in contrast to KPT, this cost depends both on the race of the officer and the motorist. This is also the key to identify racial prejudice in this model. The benefit of an arrest is one and the cost of a search is a fraction of the benefit. Without loss of generality, guilty motorists are always caught.

AF introduce definitions of prejudice and monolithic behavior. The police are prejudiced if $t(B, r_p) \neq t(W, r_p)$. In other words, the police are prejudiced if for any

[3] The model can be generalized to incorporate rational choice by the motorists

given officer race, the search costs depend on the race of the motorist. In contrast, the police exhibit monolithic behavior if $t(r_m, B) = t(r_m, W)$ for all r_m, that is to say, if all officer races have the same search costs against a given race of motorists. Conversely, the police exhibit nonmonolithic behavior if the officer groups have different search costs against a given race of motorists. It is crucial to differentiate between prejudice and monolithic behavior. A nonmonolithic police force does not necessarily indicate the presence of taste-based discrimination. It may be that one group of officers has higher search costs in general, no matter the race of the motorists. Likewise, a monolithic police force does not necessarily mean that the police are unprejudiced. It could be possible that all police groups are equally prejudiced against a particular motorist group.

Officers maximize their utility by searching at an optimal rate. Let the utility of not searching be zero. An officer will search a motorist if and only if

$$\Pr(G \mid r_m, \theta) \geq t\left(r_m, r_p\right) \tag{4}$$

where $\Pr(G \mid r_m, \theta)$ indicates the guilt probability of an r_m motorist with signal θ. This yields the minimal signal intensity required to make a search worthwhile. A police officer will only search if the signal is indicative enough of criminal behavior, so if any only if

$$\theta \geq \theta^*\left(r_m, r_p\right)$$

where the threshold θ^* is specified by (4). Intuitively, the threshold value θ^* is monotonically increasing in search costs: For a search to remain profitable when search costs are high, an officer will require a large likelihood of criminal activity (as reflected by a high θ). The threshold value θ^* directly determines the equilibrium search rate $\gamma\left(r_m, r_p\right)$ and the equilibrium search success rate (or hit rate) $S\left(r_m, r_p\right)$.

Given this setup, the rankings of the search rates and the search success rates follow a systematic pattern that is directly linked to θ^*. Assume that the search costs against motorist group B are lower than they are against motorist group W. As in KPT, this implies racial animus against group B. An officer with lower search costs is thus more likely to search group B because, in contrast to W-motorists, some otherwise unsuspicious B-motorists now become eligible for search. However, among the larger number of searched B-motorists, a lower fraction is actually guilty because relatively more innocent ones are searched. AF make use of this inverse relationship.

The model gives rise to two testable implications. First, in a monolithic police force the search costs against a given race of motorists is the same no matter the race of the officer. Therefore, all officers apply the same search criterion θ^* against that motorist race such that all search (success) rates against that motorist race are equal. This constitutes the test for monolithic behavior.

The second testable implication of AF's model exploits the inverse rank pattern of the search (success) rates in case of a nonmonolithic police force. In the absence of racial animus, the *rankings* of the search and search success rates will not depend on motorist race. Any empirical differences in search (success) rates are explained by the fact that the officer groups exhibit distinct search costs.

To illustrate, let black officers have higher search costs against white motorists than white officers do. In the absence of prejudice, this implies that the search costs of black officers against black motorists are the same as they are against white motorists. Likewise, the search costs of white officers against black motorists are the same as they are against white motorists. It then follows that the search costs of black officers against black motorists must also be higher than the search costs of white officers against black motorists. In other words, black officers have *higher search costs in general* which are independent of the race of the motorist. Formally speaking, the ranking of the search costs by officer race does not depend on the race of the motorist. Since the search and search success rates are strictly monotone functions of the search costs, the same independence holds true for the search (success) rates. Black officers will have lower search rates than white officers against any motorist race (because the higher search costs imply a higher signal threshold). On the other hand, black officers will have higher search success rates than white officers against any motorist race (because among the few searched motorists, a higher fraction is actually guilty). It takes a lot for black officers to search. But if they do, they are very successful because they focus on the motorists with the highest guilt probabilities.

The second testable implication addresses this rank independence. Lower costs against the discriminated group tend to violate the independent rank order. So the police are said to be racially prejudiced if the rankings of the search (success) rates depend on motorist race. It remains unclear, however, which officer race(s) are actually prejudiced. The test can only assess relative racial prejudice in the police force because the unbiased rank order is unknown. AF test their model on traffic data from the Florida Highway Patrol from 2000 to 2001 and cannot reject the null hypothesis of no racial prejudice.

Like the KPT test, the test proposed in AF belongs to the class of outcome tests. It assumes that police officers make best use of the observable information (r_m, θ) when deciding whether to invest in a search. Equivalently to KPT, taste based discrimination is modeled via lower search costs. But in contrast to KPT, AF's test does not need the assumption of rational choice by the motorists in order to address the infra-marginality problem. Instead, they circumvent the issue via ordinal conditions of the search (success) rates. A cardinal comparison of the rates is not valid. But because of the monotone likelihood ratio property, marginal changes affect the average outcome in a unique direction. Thus the ordinal comparison of the rates provides a valid detour around the infra-marginality problem.

AF's test shares one drawback with KPT's. It is unclear whether an unbiased police force make actual use of race in their guilt assessment. Some characteristics that give rise to a higher θ may simply be more prevalent in one group. If so, race merely correlates with guilt rates. On the other hand, race may indeed be beneficial for inferences of guilt. If black motorists have a higher propensity of crime ($\pi^{r_b} > \pi^{r_w}$), the police do best to combine this information with the individual signal θ. The following equation highlights this point in AF's model:

$$\Pr(G \mid r_m, \theta) = \frac{\pi^{r_m} f_g^{r_m}(\theta)}{\pi^{r_m} f_g^{r_m}(\theta) + (1 - \pi^{r_m}) f_n^{r_m}(\theta)} \tag{5}$$

The (posterior) probability of guilt given θ and observable race is derived via Bayes' rule (AF, p. 135). This probability can increase for the two reasons described above. First, signals indicative of crime might appear with higher probability in one group. Second, the propensity of crime might be higher in one group. Only in the latter case do the police use race in their assessment. Like in all models of statistical discrimination, this is the case when $\theta^*(B, r_p) \neq \theta^*(W, r_p)$, that is to say, when the police apply different signal thresholds for searching black and white motorists (despite being unbiased).

Equation (5) mirrors the pivotal insight of models of statistical discrimination: The assessment about a particular individual with observable signal θ depends on the distribution of that signal in the individual's group. In both KPT's and AF's tests, however, the researcher can only make inferences about whether the police have equalized the *probabilities of guilt* and not the *individual suspicion thresholds* (as reflected by the signals) along motorist race. Both modeled outcome tests inherit this feature from the generic outcome test methodology. Section 2.4 stressed that outcome tests implicitly assume that the decision-maker exploits statistical discrimination if it improves the assessment. Hence, because the probabilities of guilt may or may not include inferences based on race, one cannot be sure whether race played an instrumental role in the profile. Like generic outcome tests, the models presented in this section have nothing to say about this.

4 Legal background

4.1 United States

In US courts, statistical evidence is often used in racial profiling litigation. In *New Jersey v. Pedro Soto*, for example, a statistical report analyzed the fractions of minorities among New Jersey Turnpike motorists which had been stopped and potentially searched between 1988 and 1991 (Lamberth 1994).[4] Not surprisingly, minority motorists were at a much higher risk of being stopped or searched despite a similar probability of finding engagement in criminal activity. In the report, the null hypothesis of equal stop or search rates along race was rejected with high statistical significance. But disparate impact alone does not hold up in US courts.

At first glance, it would seem intuitive that litigation should be based on the Fourth Amendment of the US Constitution which guards against unreasonable searches. The current legal approach in racial profiling litigation however is codified in the so-called McCleskey standard and focuses on the Equal Protection Clause (EPC), which is part of the Fourteenth Amendment (Persico and Castleman 2005).[5] In addition to disparate impact, the EPC requires intent to discriminate for

[4] New Jersey v. Pedro Soto, 734 A.2d 350 (1996)

[5] McCleskey v. Kemp, 481 U.S. 279 (1987)

successful litigation. In economic terms, the plaintiff has to prove that the police exhibited taste-based discrimination in order for jurisprudence to apply strict scrutiny to the case.

The previous sections have shown that it is not an easy task to infer motives from outcome data. Disparate impact could also be the result of statistical discrimination, a goal-oriented policing strategy. And of course, race may also simply correlate with characteristics indicative of criminal behavior. This latter possibility, however, is not the relevant distinction in US jurisdiction. Baker and Phillipson (2011) describe the US approach as follows: "If a measure distinguishes on the basis of gender or illegitimacy, it must substantially advance an important state interest, and if it classifies on the basis of race, national origin, or alienage, it must be narrowly tailored to a compelling state interest." (p. 111) Disparate impact alone does not constitute a valid argument. If the state interest—be it prevention of drug trafficking or terrorist attacks—is compelling and if there are no more efficient ways to pursue this interest, the measure complies with the Constitution, no matter the degree of disparate impact. Distinguishing prejudice from efficient policing, then, is the key element in racial profiling litigation in the US (see Persico 2006 for an extensive legal review).

The statistical report by Lamberth in *New Jersey v. Pedro Soto* is not suited to differentiate between these two motives. The analysis neglects any confounding variables that influence stop and search decisions. The previous sections have suggested that an officer's assessment of a motorist's likelihood of criminal engagement can be modeled via a signal that incorporates relevant characteristics such as the condition of the car, driving behavior, or current location. Once stopped, other characteristics complete the signal: age, sex, or race of the motorist, conduct, or clearer cues clues like smell or evidence in visual range. In statistical terms, Lamberth's report does not address omitted variable bias.

Newer traffic stop data include some of this omitted information. Close and Mason (2007) make use of such data and test for racial prejudice with a logistic regression. They regress enforcement action to motorist and officer characteristics and circumstantial data such as poverty and crime rates at the stop location. Their analysis reveals that despite holding these factors constant, race remains a highly significant predictor for enforcement action. Even so, omitted variable bias remains an issue. This is where the modeled outcome tests step in.

The models exposited in Sect. 3 are therefore not theoretical gimmicks. They provide testable implications that are consistent with the current legal approach in the United States. For instance, *Anderson v. Cornejo* makes an explicit distinction between search and search success rates that substantiates the practical application. According to Judge Easterbrook, disparities in search rates do not establish an inference of prejudice. Instead, he stresses that equal search success rates "show that Custom officials search black women with (on average) the same degree of suspicion that leads them to search white women or white men."[6] Note that this coincides with the interpretation of efficient policing behavior in the modeled outcome tests. In an unbiased environment, police officers equalize the probabilities

[6] Anderson v. Cornejo, 335 F3d 1024-25 (7th Cir 2004)

of guilt between motorist races. The fact that some groups suffer from disparate impact is not relevant to the legal discussion in the United States.

4.2 Europe

Things look different in Europe. Racial profiling litigation has no pronounced history in European courts, and the topic has only recently begun to stir up public controversy. Accordingly, the legal literature is scarce (see Baker (2007) for an exception), and economic literature has not yet shown how the existing methodology pertains to Europe.

Some differences in the basic legal building blocks are evident, however. US and European law as defined by Article 14 in the European Convention on Human Rights (ECHR) differ in one dimension which qualifies the existing economic contribution to the legal discourse on racial profiling. While racial prejudice would equally hold up in court, European jurisdiction also takes issue with disparate impact. As laid out in Baker (2007) and Baker and Phillipson (2011), a plaintiff arguing on basis of the ECHR as applied, say, under the UK Human Rights Act, does not need proof of intent to discriminate for litigation. It suffices to show disparate impact. There is not even need to present evidence that the state measure in question employed statistical, let alone taste-based discrimination.

Litigation under the ECHR would, however, take into account the *degree* of disparate impact when evaluating the policing measure. This degree of negative impact is juxtaposed with the legitimate state interest the responsible measure pursues. In contrast to the US, the jurisprudence is therefore not absolute but rather a balancing act of state interest and cost to society. There can only be justification for disparate impact if so-called proportionality allows for it, that is to say, if the state interest outweighs the disparate impact. In this cost/benefit analysis, motive plays a crucial role for it affects whether the disparate impact is considered justified (Baker and Phillipson 2011, pp. 112–113).

It goes without saying that malevolent motives violate Article 14 of the ECHR. To this extent, the modeled outcome tests in this paper are of use for racial profiling litigation under the ECHR. But the evaluation of proportionality becomes more intricate when non-malignant motives are of interest. It is plausible that measures which overtly encourage the use of statistical discrimination receive higher scrutiny than measures that focus on characteristics that correlate with ethnicity. In the first case, suspects with the same individual signals are treated differently because of their ethnicity. In the second case, suspects with the same individual signals are treated equally. While this distinction is irrelevant in US litigation, it is crucial for Europe.

This paper has laid out why the existing modeled outcome tests cannot disentangle these two distinct causes. Recall that unbiased officers equalize the probabilities of suspicion along ethnicity. Equation (5) implies that these probabilities rely on group information if and only if it improves the statistical assessment. So if group information turns out to be an informative trait, the models implicitly assume that this information is taken into account. But they cannot assess the actual occurrence of such statistical discrimination, i.e. whether suspects would

have been treated differently if their ethnicity had not been observable, precisely the central question for proportionality in racial profiling litigation under the ECHR.

5 Conclusion

Racial profiling litigation in the US and in Europe relies on the identification of motives causing disparate impact. In the US, where disparate impact alone does not build a successful case, the current legal approach requires evidence of racial prejudice. On the other hand, US law has no issues with considerations of race or gender for instrumental profiling purposes, even if this entails disparate impact. Incidentally, this distinction corresponds to the economic notions of taste-based and statistical discrimination. Recently, these notions have been incorporated in modeled outcome tests, which lend themselves to making this particular distinction, a distinction that is pertinent to US discrimination law.

The legal approach in Europe, however, takes issue with disparate impact itself. This inequality can only be justified in light of a compelling state interest. More precisely, disparate impact must be evaluated with respect to motive. While racial prejudice is unambiguously shunned, statistical discrimination gives rise to higher scrutiny than mere correlation of race with characteristics indicative of criminal engagement. This paper lays out why (modeled) outcome tests are not able to make the latter distinction between what can be described as color-indifferent and color-blind. To this extent, the tests are of limited use in European jurisdiction.

In theory, there is a simple way to establish whether the police were actually color-blind in a given situation. If outcomes do not change when, ceteris paribus, ethnicity becomes invisible, ethnicity held no relevant information and was ignored in the assessment. Any disparate impact is then necessarily by correlation. Likewise, if outcomes do change when ethnicity becomes invisible, the police were considering it as a factor in their decision-making. Whether the consideration in the visible case was taste-based or statistical can be deduced by (modeled) outcome tests.

Unfortunately, this approach seems rarely viable in practice. In most settings, it is not obvious how one would exclusively render ethnicity invisible, holding everything else constant. Such experiments are generally only practical if the emitted signals of interest are not immediately tied to the body, for example in orchestra auditions behind a veil or in audit studies like written job applications. One notable exception with non-experimental data such as motor vehicle searches is presented in a study by Grogger and Ridgeway (2006). The authors exploit natural changes in daylight that affect an officer's decision to stop a motorist and argue that race is not equally well visible at day and at night. Consequently, any according variation in stopping behavior should be attributable to race. Their results show that the fraction of black motorists stopped by the police hardly varies with time of day. This suggests that in their data, race plays no role when making a stop decision.

Disentangling statistical discrimination from correlation is not only of interest for European racial profiling litigation. It also informs policy design that aims to prevent disparate impact. In motor vehicle searches, for instance, Persico (2002) shows that enforcing fairness in form of equalizing search rates does not necessarily

clash with economic efficiency (defined by minimizing the aggregate crime rate).[7] But any remedial policy should take into account that increasing fairness, say, by forcing the police to ignore race will be mostly ineffective if the main cause of disparate impact is correlation rather than statistical discrimination.

All the same, prohibiting the use of statistical discrimination inevitably comes at a price. Outcome tests require unrestricted use of statistical discrimination if it improves the assessment. The very introduction of remedial measures that restrict its use would therefore inhibit inferences of racial prejudice based on outcome tests.

Acknowledgments I would like to thank an anonymous referee for helpful comments in improving the manuscript.

References

Aigner, D. J., & Cain, G. G. (1977). Statistical theories of discrimination. *Industrial and Labor Relations Review*, *30*(2), 175–187.

Alesina, A., & La Ferrara, E. (2011). *A test of racial bias in capital sentencing*. Harvard Institute of Economic Research, Discussion Paper No 2192.

Altonji, J. G., & Blank, R. M. (1999). *Handbook of labor economics*, North Holland, chap 48: Race and Gender in the Labor Market, pp. 3143–3259.

Antonovics, K. L., & Knight, B. G. (2004). *A new look at racial profiling: Evidence from the boston police department*. NBER Working Paper 10634.

Anwar, S., & Fang, H. (2006). An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *American Economic Review*, *96*(1), 127–151

Anwar, S., & Fang, H. (2012). *Testing for racial prejudice in the parole board release process: Theory and evidence*. NBER Working Paper 18239.

Anwar, S., & Fang, H. (2013). Testing for the role of prejudice in emergency departments using bounceback rates. *BE Journal of Economic Analysis & Policy* (forthcoming).

Arrow, K. (1972). Some mathematical models of race in the labor market. In: Pascal, A. H. (Ed.), *Racial Discrimination in Economic Life*, Lexington, MA: Lexington Books.

Ayres, I. (2001). *Pervasive Prejudice? Unconventional evidence of race and gender discrimination*. Chicaco: The University of Chicaco Press.

Ayres, I. (2002). Outcome tests of racial disparities in police practices. *Justice Research and Policy*, *4*, 131–142.

Ayres, I. (2005). Three tests for measuring unjustified disparate impacts in organ transplantation. *Perspectives in Biology and Medicine*, 48(1, supplement):S68–S87.

Ayres, I., & Waldfogel, J. (1994). A market test for race discrimination in bail setting. *Stanford Law Review*, *46*(5), 987–1047.

Baker, A. (2007). Controlling racial and religious profiling: Article 14 echr protection v. u.s. equal protection clause prosecution. *Texas Wesleyan Law Review*, *13*, 285–309.

Baker, A., & Phillipson, G. (2011). Policing, profiling and discrimination law: US and European approaches compared. *Journal of Global Ethics*, *7*(1), 105–124.

Becker, G. S. (1957). *The economics of discrimination*. Chicago: University of Chicago Press

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, *76*(2), 169–217.

Becker, G. S. (1993). *The evidence against banks doesn't prove bias*. Business Week April 19th:18

Close, B. R., & Mason, P. L., (2007). Searching for efficient enforcement: Officer characteristics and racially biased policing. *Review of Law and Economics*, *3*(2), 263–321.

Dharmapala, D., & Ross, S.L., (2004). Racial bias in motor vehicle searches: Additional theory and evidence. *Contributions to Economic Analysis & Policy*, *3*(1): Article 12

[7] The reason for this proposition lies in the divergence of the interest of the police on the one hand and the state on the other. In most models, the objective of a police officer is the maximization of the search success rate, which is typically not in accordance with the goal of minimizing the aggregate crime rate.

Dodd, V. (2005). *Asian men targeted in stop and search*. The Guardian 17 August 2005.
Fang, H., & Moro, A. (2011). *Theories of statistical discrimination and affirmative action: A survey*,Vol. 1A, North, chapter 5, pp 133–200.
Grogger, J., & Ridgeway, G. (2006). Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association*, *101*(475), 878–887.
Heckman, J.J. (1998). Detecting discrimination. *Journal of Economic Perspectives*, *12*(2), 101–116.
Ilić, D. (2013). Spatial and temporal aggregation in racial profiling. *Swiss Journal of Economics and Statistics*, *149*(1), 27–56.
Knowles, J., Persico, N., & Todd, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, *109*(1), 203–229.
Lamberth, J. (1994). Revised statistical analysis of the incidence of police stops and arrests of black drivers/travellers on the New Jersey Turnpike between interchanges 1 and 2 from the years 1988 through 1991, legal Expert Report
Lever, A. (2005). Why racial profiling is hard to justify: A response to risse and zeckhauser. *Philisophy & Public Affairs*, *33*(1), 94–110.
Loury, G. C. (2002). *The anatomy of racial inequality*. Harvard: Harvard University Press.
Loury, G.C., & Coate, S. (1993). Will affirmative action policies eliminate negative stereotypes? *American Economic Review*, *83*(5), 1220–1240.
Persico, N. (2002). Racial profiling, fairness, and effectiveness of policing. *American Economic Review*, *92*(5), 1472–1497.
Persico, N. (2006). Rational choice foundations of equal protection in selective enforcement: Theory and evidence, u of Penn, Inst for Law & Econ Research Paper No. 06–20.
Persico, N. (2009). Racial profiling? Detecting bias using statistical evidence. *Annual Review of Economics*, *1*, 229–254.
Persico, N., & Castleman, D.A. (2005). Detecting bias: Using statistical evidence to establish intentional discrimination in racial profiling cases. *University of Chicago Legal Forum*, *1*, 1–19.
Persico, N., & Todd, P. (2005). Passenger profiling, imperfect screening and airport security. *American Economic Review*, *95*(2), 127–131.
Persico, N., & Todd, P. (2006). Generalising the hit rates tests to test for racial bias in law enforcement, with an application to vehicle searches in wichita. *The Economic Journal*, *116*, F351–F367.
Persico, N., & Todd, P.E. (2008). The hit rates test for racial bias in motor vehicle searches. *Justice Quarterly*, *25*(1), 37–53.
Phelps, E. (1972). The statistical theory of racism and sexism. *American Economic Review*, *62*, 659–661.
Risse, M., & Zeckhauser, R. (2004). Racial profiling. *Philisophy & Public Affairs*, *32*(2), 130–171.
Smart, S., Shoven, J., & Waldfogel, J. (1996). *A citation-based test for discrimination at economic and finance journals*. NBER Working Paper 5460.
Yinger, J. (1996). Why default rates cannot shed light on mortgage discrimination. Cityscape: *A Journal on Policy Development and Research*, *2*(1), 25–31.
Yinger, J. (1998). Evidence on discrimination consumer markets. *Journal of Economic Perspectives*, *12*(2), 23–40.